APPENDIX D

CONSTRUCTION OF SAMPLE WEIGHTS AND STANDARD ERRORS FOR ANALYSES USING BASELINE INTERVIEW AND PROGRAM INTAKE DATA

THIS PAGE INTENTIONALLY BLANK

A. INTRODUCTION

This technical appendix describes the construction of sample weights so that statistics based on baseline interview and sample intake data can be generalized to the study population for the National Job Corps Study. In addition, it discusses procedures used to construct standard errors of the estimates.

B. CONSTRUCTION OF SAMPLE WEIGHTS

Youths in the study population had different probabilities of being assigned to the program research and control groups, because sampling probabilities differed for various population subgroups. In addition, youths in the research sample had different probabilities of being included in the baseline interview sample, because (1) baseline interview attempts continued in the post-45-day period for sample members who lived in randomly selected areas only, and (2) youths in different types of areas (superdense, dense, and nondense) had different probabilities of being eligible for post-45-day baseline interviews.

Next, we discuss how weights were constructed to account for these design features. We conclude the section with a discussion of our approach for adjusting the weights to account for the effects of nonresponse to the baseline interview.

1. Sample Design Weights

The sample design weight for a sample member was constructed to be proportional to the inverse of the probability that the youth was selected into the research sample. Table D.1 displays selection probabilities by research status for individuals in those subgroups for which sampling rates were constant. The sampling rates to the control group are displayed by gender and whether the

TABLE D.1

PROBABILITIES THAT ELIGIBLE APPLICANTS WERE SELECTED TO THE CONTROL AND PROGRAM RESEARCH GROUPS, BY SAMPLING STRATA
(Percentages)

	Sampling Probability	
	Random Assignment Date Before 8/16/95	Random Assignment Date on or After 8/16/95
Control Group		
Females in areas from which a low concentration of nonresidential Job Corps female students come	5	5
Females in 57 areas from which a high concentration of nonresidential Job Corps female students come	8	9
Males in areas from which a low concentration of nonresidential Job Corps female students come	8	8
Males in 57 areas from which a high concentration of nonresidential Job Corps female students come	8	9
Program Research Group		
Residential designees	10.7	11.1
Nonresidential designees	15.4	17.0
Number in Sample Universe	**47,288**	**33,595**

youth lived in the 57 areas sending the largest number of nonresidential students to Job Corps.[1] The sampling rates to the program research group are displayed by residential designation status obtained from the ETA-652 Supplement. The control and program research group sampling rates are displayed also for youths who were sent for random assignment before and after August 16, 1995. This is because the probabilities that youths were assigned to the research sample were increased for likely nonresidential students at that time to compensate for the lower-than-expected flow of eligible applicants and the higher-than-expected program no-show rate.

The sampling probabilities displayed in Table D.1 were adjusted for the following sample members:

- Four youths in the program research group who were also randomly assigned to the program nonresearch group.[2] The selection probabilities for each of these youths is *2p* where *p* is the relevant sampling probability from Appendix Table D.1 for each youth.

- 27 youths who were recruited by the Florida employment service office in Hialeah (FLESHI) and who were randomized to the research sample after March 27, 1995. A large proportion of youths recruited by FLESHI in early 1995 were assigned to the control group, and FLESHI staff expressed concern to Region 4 senior staff about the negative effects the evaluation was having on their reputation. To help smooth the flow of control group members who were recruited by FLESHI for the remainder of the sample intake period, all youths sent for random assignment after March 27, 1995, had the *same* probability of being assigned to the control group (and the same probability of being assigned to the program research group). Hence, all youths in a batch sent for random assignment were randomized together rather than in separate strata. The uniform sampling rates were set as the average of all the sampling probabilites of all FLESHI youths who were sent for random assignment prior to March 28, 1995. The sampling rates to the control group were set as follows: (1) 7.63 percent for those sent for random assignment between March 28, 1995, and August 15, 1995; and (2) 8.05 percent for those sent for random assignment after August 15, 1995. The sampling rates to the program research group were set as follows: (1) 11.62 percent for those sent for

[1]Sampling rates were higher in these 57 areas to meet sample size targets for nonresidential students.

[2]This occurred as the result of an error in our random assignment program, which did not check whether duplicate information on a youth was present *within* a batch of information sent to MPR for random assignment purposes.

random assignment between March 28, 1995, and August 15, 1995; and (2) 12.04 percent for those sent for random assignment after August 15, 1995.

Sample design weights were constructed by first calculating the inverse of the selection probabilities and then scaling the resulting weights so that they sum to 5,977 for control group members and 9,409 for program research group members (which are the sample sizes of the control and program research groups). These weights were applied in analyses using ETA-652 and Supplemental ETA-652 data.

2. Baseline Interview Weights

As discussed in detail in Appendix C, baseline interviews were attempted by telephone with all youths in the research sample during the first 45 days after random assignment. However, only youths in randomly selected areas who were not reachable by telephone within the 45-day period were eligible for telephone or in-person interviews during the post-45-day period.[3] To select these areas, we divided the country into 16 superdense, 29 dense, and 75 nondense areas. We then selected all 16 superdense, 18 dense, and 29 nondense areas as those where youths would be eligible for post-45-day interviewing. We selected different proportions of superdense, dense, and nondense areas for in-person interviewing to maximize the precision of the impact estimates, subject to the cost of conducting interviews in each type of area and a fixed interview budget.

The within-45-day sample is a random sample of those in the study population reachable by telephone within 45 days. The post-45-day sample, however, is a *clustered* sample of those in the study population reachable by telephone after 45 days. Thus, the post-45-day sample is

[3]Control group members designated for nonresidential slots on the Supplemental ETA-652 form, however, were eligible for post-45-day interviews regardless of where they lived. This design feature was adopted to increase the precision of impact estimates for the small nonresidential program component.

underrepresented in the baseline sample relative to their numbers in the study population, and those in superdense, dense, and nondense areas have different representations in the post-45-day sample.

For analyses using baseline interview data, the weight for a youth--the interview weight--was constructed to be proportional to the inverse of the probability that the youth was selected into the baseline interview sample. This probability was calculated by multiplying the probability the youth was selected into the research sample (as described above) by a factor *f* defined as follows:

$f = 1$	if the youth completed a baseline interview within the first 45 days after random assignment
$= 1$	if the youth lived in a superdense area at application to Job Corps
$= 1$	if the youth was in the control group and was designated for a nonresidential slot on the Supplemental ETA-652 form
$= 18/29$	if the youth completed a baseline interview after the 45-day period and lived in a dense area at application to Job Corps
$= 29/75$	if the youth completed a baseline interview after the 45-day period and lived in a nondense area at application to Job Corps

The factor *f* can be interpreted as the conditional probability that an eligible applicant was in the baseline sample given that the individual was selected into the research sample. The interview weights pertaining to the baseline interview were scaled to sum to 5,514 for control group members and to 8,813 for program research group members (which are the number of control and program research group members who completed baseline interviews).

It is important to note that the overall weighted mean of a survey data item can be computed as follows:

(1) $\bar{y} = J\bar{y}_1 + (1-J)[2_s\bar{y}_{2s} + 2_d\bar{y}_{2d} + 2_n\bar{y}_{2n}]$,

where:

$\bar{y}$ = the overall weighted mean of the variable

$\bar{y}_1$ = the weighted mean (using the sample design weights) of those in the sample who completed baseline interviews within 45 days after random assignment

$\bar{y}_{2s}$, $\bar{y}_{2d}$, $\bar{y}_{2n}$
= the weighted mean (using the sample design weights) of those who completed baseline interviews in the post-45-day period in superdense, dense, and nondense areas, respectively

2_s, 2_d, 2_n
= the proportion of the post-45-day population in superdense, dense, and nondense areas, respectively

J = the proportion of all potential baseline interview completers who would complete the baseline interview within 45 days after random assignment

The procedure we use to construct the interview weights assumes that the weight, J, is the proportion of baseline interview completers in the selected in-person areas who completed the baseline interview within 45 days (which is about 93 percent). This assumes that baseline interview nonrespondents are split *proportionally* between the within-45-day and post-45-day populations. As discussed next, this is probably a reasonable assumption because the characteristics at program intake of baseline interview nonrespondents, within-45-day responders, and post-45-day responders are similar.

3. Adjustments for Nonresponse

The effective response rate to the baseline interview was over 95 percent. However, descriptive statistics estimated using baseline interview data could be slightly biased if the characteristics of interview respondents and nonrespondents differ. In this section, we assess the effects of baseline

nonresponse and discuss our approach for adjusting for these effects. First, we discuss the data and methods used in the analysis. Second, we discuss analysis results.

a. Data and Methods

Our basic approach for assessing the effects of nonresponse is to compare the characteristics of respondents and nonrespondents by using ETA-652 and ETA-652 Supplement data, which were collected at program intake and thus are available for both interview respondents and nonrespondents. For the analysis, we select data items that we believe are correlated with (1) whether a youth was a respondent, and (2) key baseline measures and outcomes.

The analysis is performed using *only* those sample members who lived in the areas selected for post-45-day followup at application to Job Corps. Youths in the nonselected areas are excluded from the analysis, because "nonrespondents" in these areas consist of both those who would and those who would not have completed interviews in the post-45-day period if given the chance. Therefore, "true" nonrespondents can be identified only in the selected areas. This sample of nonrespondents, however, is representative of nonrespondents nationwide. The analysis sample contains 10,026 respondents (4,037 control group and 5,989 program research group members) and 514 nonrespondents (249 control group and 265 program research group members).

As part of the analysis, we compare respondents in the in-person areas who completed the interview within 45 days after random assignment and those who completed the interview after 45 days. We also compare these two groups to nonrespondents. This analysis is needed to assess how statistics computed using the within-45-day and post-45-day samples should be weighted to produce overall statistics. For example, if interview nonrespondents are more similar to those in the post-45-day sample than to those in the within-45-day sample, then the statistics using the post-45-day

sample should be given a weight larger than the proportion of interview respondents in the in-person areas who completed interviews during the post-45-day period (see Section B.3.b).

We use standard statistical tests to assess the similarity of respondents and nonrespondents and of within-45-day and post-45-day respondents. We use univariate t-tests to compare variable means for binary and continuous variables and chi-squared tests to compare variable distributions for categorical variables. In addition, we conduct a more formal multivariate analysis to test the hypothesis that key variable means and distributions are *jointly* similar. For this analysis, we estimate logit regression models where the probability an individual is a respondent versus a nonrespondent is regressed on a set of youth characteristics. Chi-squared (log-likelihood) tests are used to assess whether the explanatory variables in the models are jointly statistically significant.

b. Analysis Results

There are some differences in the characteristics of baseline interview respondents and nonrespondents (see Table D.2). Younger sample members were more likely than older sample members to complete a baseline interview. In addition, response rates were higher (1) for youths who did not need a bilingual Job Corps program than for those who did, (2) for those who lived in large families than for those who lived in smaller families, (3) for those without criminal backgrounds than for those with criminal backgrounds, and (4) for those who applied to Job Corps earlier than for those who applied later. There are, however, few significant differences in the other variables between the two groups. The distributions of respondents and nonrespondents are similar by gender, race, region, size of city, PMSA or MSA residency status, the presence of dependents, education level, the receipt of welfare, and anticipated program enrollment variables. There are few

TABLE D.2

COMPARISON OF THE CHARACTERISTICS OF RESPONDENTS AND NONRESPONDENTS TO THE BASELINE INTERVIEW, BY RESEARCH STATUS
(Percentages)

	Control Group		Program Research Group	
	Respondents	Nonrespondents	Respondents	Nonrespondents
Demographics				
Male	58.0	58.3	58.0	63.2*
Age at Application				
16 to 17	40.2	29.9***	40.2	29.7***
18 to 19	31.8	35.3	32.2	29.2
20 to 21	17.0	14.5	16.1	21.3
22 to 24	11.0	20.3	11.5	19.8
(Average age)	18.9	19.6***	18.9	19.7***
Race/Ethnicity				
White, non-Hispanic	24.1	24.8***	24.3	22.7*
Black, non-Hispanic	54.8	50.8	55.6	54.4
Hispanic	16.7	15.0	15.7	14.6
American Indian or Alaskan Native	2.3	3.4	2.0	3.7
Asian or Pacific Islander	2.1	6.0	2.4	4.6
Job Corps Region of Residence				
1	5.4	5.7	5.3	6.3**
2	9.2	10.5	9.0	5.8
3	14.7	10.8	13.7	14.3
4	21.5	19.1	22.3	22.1
5	9.7	10.4	9.6	16.5
6	13.2	17.2	14.3	10.4
7/8	11.0	13.1	11.9	10.2
9	10.3	9.6	9.2	8.4
10	5.1	3.7	4.6	5.9
Size of City of Residence				
Less than 2,500	5.4	5.6	5.4	4.7
2,500 to 10,000	7.2	5.4	8.0	5.7
10,000 to 50,000	15.1	14.4	15.7	17.5
50,000 to 250,000	18.1	16.8	18.3	19.6
250,000 or more	54.1	57.8	52.6	52.5
PMSA or MSA Residence Status				
In PMSA	44.0	49.1	45.1	43.9
In MSA	43.1	40.1	41.6	45.4
In neither	12.9	10.8	13.3	10.8
Type of Area				
Superdense	49.9	49.6	51.4	48.1
Dense	26.7	26.1	25.3	27.8
Nondense	23.4	24.3	23.4	24.1
In 57 Areas Sending the Largest Number of Nonresidential Females to Job Corps	40.1	40.4	37.3	39.0
Legal Resident	98.8	99.1	98.5	98.9
Needs Bilingual Program in Job Corps	4.2	8.0***	3.9	7.7***

TABLE D.2 *(continued)*

	Control Group		Program Research Group	
	Respondents	Nonrespondents	Respondents	Nonrespondents
Job Corps Application Date				
11/94 to 2/95	22.0	18.4*	23.2	15.5**
3/95 to 6/95	30.0	28.6	28.7	33.0
7/95 to 9/95	28.4	26.6	27.8	31.4
10/95 to 12/95	19.6	26.4	20.3	20.1
Fertility and Household Composition				
Has Dependents	16.9	16.7	15.1	14.4
Family Status				
Family head	14.2	15.5***	13.6	19.5***
Family member	61.8	49.1	61.5	47.1
Unrelated individual	24.0	35.5	24.8	33.3
Average Family Size	3.2	2.6***	3.2	2.7***
Education				
Highest Grade Completed				
Below 9	14.4	12.0	15.4	13.4
9 to 11	63.7	68.7	63.7	63.9
12	21.2	18.7	20.1	22.4
Above 12	0.8	0.6	0.8	0.4
(Average grade)	10.1	10.1	10.0	10.1
Welfare Dependence				
Type of Welfare Received				
AFDC	28.1	27.3	28.0	27.8
Other types	14.5	16.5	15.3	14.3
None	57.4	56.2	56.7	57.9
Health				
Ever Had Any Serious Illnesses or Injuries	2.1	5.0***	2.9	3.7
Have Any Health Conditions That Are Being Treated	3.1	4.6	3.5	3.5
Crime				
Arrested in Past Three Years, Other than for Minor Traffic Violations	11.5	15.2*	11.5	14.7
Ever Convicted or Adjudged Delinquent	5.4	11.0***	5.7	7.2

TABLE D.2 *(continued)*

	Control Group		Program Research Group	
	Respondents	Nonrespondents	Respondents	Nonrespondents
Anticipated Program Enrollment Information				
Designated for a Nonresidential Slot	19.9	15.2*	14.8	13.5
Designated for a CCC Center[a]	12.4	12.1	12.7	14.3
Designated for a Low or Medium Low Performing Center[a]	53.3	56.2	53.4	51.9
Designated for a Small or Medium Small Center[a]	63.3	59.4	62.4	65.6
Sample Size	**4,037**	**249**	**5,989**	**265**

SOURCE: Data From ETA-652 and ETA-652 Supplemental forms.

NOTE: The figures are calculated using only those sample members who lived in areas selected for in-person interviewing when they applied to Job Corps.

[a] Figures are obtained using data on OA counselor projections about the centers that youths were likely to attend.

*Difference between distributions for respondents and nonrespondents is significantly different from zero at the .10 level, two-tailed test.
**Difference between distributions for respondents and nonrespondents is significantly different from zero at the .05 level, two-tailed test.
***Difference between distributions for respondents and nonrespondents is significantly different from zero at the .01 level, two-tailed test.

differences in our findings by research status. The parameter estimates from the multivariate logit models yield similar results (not shown).[4]

Because the differences between the characteristics of respondents and nonrespondents are not large and do not differ by research status, we did not adjust for the effects of nonresponse in the final tabulations using baseline interview data. We did conduct the analysis, however, using adjusted weights to test the sensitivity of our estimates. The original weights were adjusted so that the weighted characteristics of interview respondents were similar, on average, to those of the full population of respondents and nonrespondents.[5] We found that the tabulations using the adjusted and unadjusted weights were almost identical. This occurred because response rates to the baseline interview were high so that adjusting for nonresponse had only a small effect on the overall estimates. In addition, the adjustments to the original sample weights were small, because our model could not accurately distinguish between respondents and nonrespondents on the basis of available youth characteristics.[6]

There are also some differences between the characteristics of respondents who completed the baseline interview within 45 days after random assignment and respondents who completed the interview during the post-45-day period, and the patterns are similar for program and control group

---

[4]The explanatory variables in the logit models are jointly statistically significant at the 1 percent level of significance for both program and control group members. This result, however, is caused by the statistical significance of a small subset of variables.

[5]The basic procedure we used for constructing these weights was to (1) create a predicted probability (propensity score) for each respondent and nonrespondent using estimates from the "best" logit model (which included only variables with predictive power), (2) divide the youths into six groups on the basis of the size of their predicted probabilities, and (3) calculate the (weighted) interview response rate in each group. The adjusted weight for a youth was then constructed to be inversely proportional to the product of the original weight and the response rate in that youth's group.

[6]For example, the response rate in the group with the lowest propensity scores (that is, the group with the lowest probabilities of being interview respondents) was nearly 90 percent.

members (see Table D.3). For example, there is some evidence that older sample members, those who lived in rural areas, those who needed a bilingual program, Asians, and those who lived in smaller households were more likely than their counterparts to be in the post-45-day sample. However, there are few other differences between the two groups.

While there is some evidence that the characteristics of interview nonrespondents are more similar to those of the post-45-day respondents than to those of within-45-day respondents, the differences between the three groups are not large. Consequently, our weights using baseline interview data are constructed under the assumption that interview nonrespondents are split proportionally among the two respondent groups.

C. CONSTRUCTION OF STANDARD ERRORS

The standard errors of estimates using program intake data are straightforward to calculate, although they need to account for design effects due to unequal weighting of the sample. The standard errors of estimates using baseline interview data, however, are much more complicated to calculate, because they must also account for design effects due to the clustered post-45-day sample caused by the random selection of areas for post-45-day interviewing.

In this three-part section, we discuss how we calculated standard errors for estimates based on baseline interview data. In the first section, we discuss how we estimated standard errors for a variable mean. Second, we discuss how we estimated standard errors for the difference of means across two groups. These standard errors were used to conduct t-tests to test the hypothesis that the group means are equal. Finally, we discuss how we conducted chi-squared tests to compare distributions of categorical variables across two groups.

TABLE D.3

COMPARISON OF THE CHARACTERISTICS OF BASELINE INTERVIEW RESPONDENTS WHO COMPLETED THE INTERVIEW WITHIN AND AFTER 45 DAYS AFTER RANDOM ASSIGNMENT, BY RESEARCH STATUS
(Percentages)

	Control Group		Program Research Group	
	Within-45-Day Respondents	Post-45-Day Respondents	Within-45-Day Respondents	Post-45-Day Respondents
Demographics				
Male	57.8	64.3*	58.2	58.5
Age at Application				
16 to 17	40.7	34.8	40.7	33.1***
18 to 19	31.6	35.1	32.3	33.0
20 to 21	16.9	17.8	16.0	16.5
22 to 24	10.8	12.3	11.0	17.4
(Average age)	18.9	19.2	18.9	19.3***
Race/Ethnicity				
White, non-Hispanic	24.4	23.1***	24.5	27.4***
Black, non-Hispanic	54.6	53.5	55.8	47.5
Hispanic	16.9	13.8	15.7	15.4
American Indian or Alaskan Native	2.3	4.3	2.0	3.1
Asian or Pacific Islander	1.8	5.4	2.0	6.6
Job Corps Region of Residence				
1	5.5	5.0**	5.4	4.4***
2	9.5	3.2	9.1	5.1
3	14.9	9.6	14.1	7.1
4	21.5	26.6	22.4	24.4
5	9.5	12.1	9.3	14.1
6	13.1	13.5	14.2	15.6
7/8	10.7	13.1	11.9	12.7
9	10.1	11.8	9.1	9.7
10	5.2	5.2	4.6	6.9
Size of City of Residence				
Less than 2,500	5.6	4.1***	5.4	7.5***
2,500 to 10,000	7.2	10.9	8.2	8.0
10,000 to 50,000	15.1	17.0	15.6	22.6
50,000 to 250,000	17.7	27.0	18.6	14.4
250,000 or more	54.4	41.1	52.2	47.5
PMSA or MSA Residence Status				
In PMSA	44.2	29.8***	45.1	33.9***
In MSA	42.6	54.4	41.4	47.9
In neither	13.2	15.8	13.5	18.2
Type of Area				
Superdense	49.4	39.5***	51.1	36.9***
Dense	27.1	23.4	25.3	26.8
Nondense	23.5	37.0	23.6	36.3

TABLE D.3 *(continued)*

	Control Group		Program Research Group	
	Within-45-Day Respondents	Post-45-Day Respondents	Within-45-Day Respondents	Post-45-Day Respondents
In 57 Areas Sending the Largest Number of Nonresidential Females to Job Corps	40.2	31.1***	37.3	30.4***
Legal Resident	98.8	99.5	98.5	98.8
Needs Bilingual Program in Job Corps	3.8	9.4***	3.5	9.4***
Job Corps Application Date				
11/94 to 2/95	22.1	21.1	23.6	16.3***
3/95 to 6/95	30.0	31.3	28.5	32.0
7/95 to 9/95	28.2	29.8	27.4	34.8
10/95 to 12/95	19.6	17.8	20.6	17.0
Fertility and Household Composition				
Has Dependents	16.8	16.0	15.0	15.8
Family Status				
Family head	14.0	16.6**	13.6	14.3***
Family member	62.3	52.6	62.0	53.5
Unrelated individual	23.6	30.8	24.4	32.3
Average Family Size	3.2	2.9**	3.2	3.1
Education				
Highest Grade Completed				
Below 9	14.5	10.9	15.5	12.5
9 to 11	63.6	65.4	63.7	65.1
12	21.1	22.9	20.0	21.5
Above 12	0.7	0.7	0.8	1.0
(Average grade)	10.1	10.2	10.0	10.2*
Welfare Dependence				
Type of Welfare Received				
AFDC	28.0	26.7	28.0	26.4
Other types	14.7	15.0	15.2	17.6
None	57.4	58.3	56.8	56.0
Health				
Ever Had Any Serious Illnesses or Injuries	2.1	2.1	2.9	3.0
Have Any Health Conditions That Are Being Treated	3.0	2.8	3.5	4.6

TABLE D.3 *(continued)*

	Control Group		Program Research Group	
	Within-45-Day Respondents	Post-45-Day Respondents	Within-45-Day Respondents	Post-45-Day Respondents
Crime				
Arrested in Past Three Years, Other than for Minor Traffic Violations	11.3	13.4	11.5	11.5
Ever Convicted or Adjudged Delinquent	5.4	5.7	5.6	6.5
Anticipated Program Enrollment Information				
Designated for Nonresidential Slot	20.2	11.1***	14.7	13.7
Designated for a CCC Center[a]	12.6	14.5	62.6	61.5
Designated for a Low or Medium Low Performing Center[a]	53.2	54.0	53.6	51.4
Designated for Small or Medium Small Center[a]	63.6	60.6	53.6	51.4
Sample Size	**3,785**	**252**	**5,579**	**410**

SOURCE: Data from ETA-652 and ETA-652 Supplemental forms.

NOTE: The figures are calculated using only those sample members who lived in areas selected for in-person interviewing when they applied to Job Corps.

[a] Figures are obtained using data on OA counselor projections about the centers that youths were likely to attend.

*Difference between distributions for within-45 and post-45 day respondents is significantly different from zero at the .10 level, two-tailed test.

**Difference between distributions for within-45 and post-45 day respondents is significantly different from zero at the .05 level, two-tailed test.

***Difference between distributions for within-45 and post-45 day respondents is significantly different from zero at the .01 level, two-tailed test.

1. Standard Error of a Variable Mean

The variance of a mean measure can be written using equation (1) as follows:

$$(2) \qquad var(\bar{y}) = J^2 var(\bar{y}_1) + (1-J)^2 [2_s^2 var(\bar{y}_{2s}) + 2_d^2 var(\bar{y}_{2d}) + 2_n^2 var(\bar{y}_{2n})] .$$

Next, we discuss the calculation of each of the variance components in equation (2).

The sample that completed baseline interviews within 45 days after random assignment is a random sample. Hence, the variance of a mean measure for the within-45-day sample (the first variance component) can be written as follows:

$$(3) \qquad var(\bar{y}_1) = (1-g) deffw_1 \frac{F_1^2}{n_1} ,$$

where:

- F_1^2 = variance of the measure in the within-45-day population
- g = proportion of the population that is sampled (which is assumed in all analyses to be the average sampling rates to the research sample--7.4 percent for control group members and 11.6 percent for program group members)
- n_1 = within-45-day sample size
- $deffw_1$ = design effect due to unequal sample design weights (w) (which equals $n_1 3 w^2 / (3w)^2$, and that is due to the fact that various population subgroups had different probabilities of being selected to the research sample)

An unbiased estimate of the unknown F_1^2 is calculated in the usual way, and the estimate is inserted in place of F_1^2 in equation (3).

The variance of a mean measure for the post-45-day sample in superdense areas--that is, $var(\bar{y}_{2s})$--is calculated in a similar way, because all 16 superdense areas were selected as in-person areas.

The post-45-day samples in dense and nondense areas, however, are clustered samples, because subsamples of these areas were selected for baseline followup after the 45-day period. The variance of the mean measure for the post-45-day sample in dense areas can be written as follows:

$$(4) \quad var(\bar{y}_{2d}) \;'\; deffw_{2d}\left[(1\&g)\frac{(1\&D_{2d})F^2_{2d}}{n_{2d}a_d} \;\%\; (1\&f_d)\frac{D_{2d}F^2_{2d}}{a_d}\right],$$

where:

F^2_{2d}	=	variance of the measure in dense areas in the post-45-day population
D_{2d}	=	proportion of the total variance that is between-area variance
f_d	=	proportion of the 29 dense areas selected for post-45-day baseline follow-up (18/29)
n_{2d}	=	average post-45-day sample size in the dense areas
a_d	=	number of dense areas selected for post-45-day baseline followup
$deffw_{2d}$	=	design effect due to unequal sample design weights

and where g is defined as in equation (3). The variance of a mean measure for the post-45-day sample in nondense areas is computed similarly.[7,8]

---

[7]Equation (4) corresponds to the variance of a mean under a design where subsampling occurs with units of equal size. This is a good approximation for the Job Corps design, because dense areas were constructed to have similar numbers of eligible Job Corps applicants, and similarly for nondense and superdense areas. The mean number of youths in our sample frame per dense area was 788, the median number was 775, and the 25th and 75th percentiles were 640 and 911, respectively. The mean number of youths in our sample frame per nondense area was 403, the median number was

(continued...)

In equation (4), the first term inside the brackets signifies the variance of the measure across youths within areas, while the second term inside the brackets signifies the variance of the mean measure across areas. If the mean measure varies little across areas (that is, if D is small), then the design effect due to clustering is small. On the other hand, if the proportion of the total variance that is between-area variance is large, then the design effect due to clustering is large. This can be seen by noting that the design effect due to clustering can be estimated by dividing the bracketed term in equation (4) by the variance of the mean measure for a random sample of the same size, which yields the following expression:

$$(5) \quad deff_{clus} = 1 + D\left[\frac{(1-f)}{(1-g)}n - 1\right],$$

where subscripts are dropped for notational simplicity. Hence, there is a one-to-one correspondence between the design effect and D for given sample sizes.

An unbiased estimate of the variance expression in equation (4) is as follows:

$$(6) \quad \hat{var}(\bar{y}) = deff_{w}\left[(1-f)\frac{s_b^2}{a} + f(1-g)\frac{s_w^2}{na}\right],$$

---

[7](...continued)
403, and the 25th and 75th percentiles were 309 and 477. Because the sample sizes did not differ significantly across the dense areas and the nondense areas, we did not adjust the weights using poststratification procedures or assume that subsampling occurred with units of unequal size.

[8]Equation (4) is an approximation because the actual variance of the mean is a weighted average of the clustered variances *across the control (program research) group sampling strata*, where the weight in each stratum is the squared percentage of those in the sample universe in that stratum. We use equation (4) because there are only a very small number of post-45-day youths in most of the sampling strata.

where s_b^2 is the sample variance of the mean measure between areas, s_w^2 is the (average) sample variance of the measure across youths within areas, and other subscripts are omitted for notational simplicity.[9]

A problem with using equation (6), however, is that the response rate to the baseline interview was extremely high within the first 45 days after random assignment (89 percent) and only an additional 6 percent of the research sample in the in-person areas completed baseline interviews in the post-45-day period. Hence, the post-45-day sample is small. The sample contains only 149 sample members (97 program research and 52 control group members) who lived in the 18 selected dense areas and 138 sample members (83 program and 55 control groups members) who lived in the 29 selected nondense areas. Hence, there are very few sample members in most of the selected dense and nondense areas, and there are none in several areas. Thus, the between-area and within-area variance estimates in the dense and nondense areas (that is, s_b^2 and s_w^2) would be imprecise if the post-45-day sample were used in the calculations.

To address this problem, we calculated the variance of a mean measure in the dense (and nondense) areas using the following two steps:

1. We estimated s_b^2 and s_w^2 in dense (nondense) areas using both the *within-45-day and post-45-day* samples who lived in the selected dense (nondense) areas.

2. Using the estimated variances in step (1), we calculated equation (6) using *post-45-day* sample sizes.

This procedure assumes that the between-area and within-area variance estimates are similar for the within-45-day and post-45-day populations. This assumption cannot be reliably tested, because of small post-45-day sample sizes. However, we believe that it is sufficiently accurate and that our

[9]The design effect (and, consequently, D) can be estimated by dividing equation (6) by an unbiased estimate of the variance of a simple random sample of the same size (that is, of *na* youths).

procedure will yield more reliable variance estimates than those that would be obtained using only the post-45-day samples in the calculations.

An estimate of the total variance of a mean measure (that is, of the expression in equation (2)) can then be calculated using the estimated variances for the within-45-day and post-45-day samples. Design effects are estimated by dividing this total variance estimate by an unbiased estimate of the variance of a simple random sample of the same size.

The total design effect for most measures based on the full baseline interview sample is about 1.07. Consequently, the standard errors of the measures are about 3.4 percent larger than those produced using standard statistical software.[10]

2. Standard Error of Differences in Two Means

In this report and the companion reports, we conducted several analyses where variable means based on baseline interview data were compared across two groups. For example, in Appendix B of this report, we compared the average characteristics of program and control group members. This section discusses how we obtained standard errors for these types of analyses. The approach we use to obtain standard errors for differences in mean measures is an extension of the approach we used in the previous section to obtain standard errors for variable means.

The variance of a difference in a mean measure can be written as follows:

$$(7) \quad var(\bar{I}) = J^2 var(\bar{I}_1) + (1-J)^2 [2_s^2 var(\bar{I}_{2s}) + 2_d^2 var(\bar{I}_{2d}) + 2_n^2 var(\bar{I}_{2n})],$$

[10]The design effect due to unequal baseline interview weights is 1.057. The design effect due to unequal sample design weights is 1.03.

where $\bar{I}$ represents the difference between the group means, and where the other parameters and subscripts were defined in the previous section.

Because these two samples are independent, the variance of the difference in means in the within-45-day sample is simply the sum of the variances of each of the group means. Thus, equation (3) applied separately to each of the two groups can be used to estimate this variance component. The same procedure can be used also to estimate the variance of the difference in means in the superdense areas.

The two samples in the post-45-day sample in dense or nondense areas, however, may not be independent, because these samples were selected from the *same* areas. For example, the average characteristics of program research and control group members who live in the same areas may be correlated, because they face similar local economic conditions and because individuals with similar characteristics tend to cluster in the same geographic areas. Thus, the average measures for the two groups in the same area may be correlated.

The variance of the difference in means in dense or nondense areas can be written as follows:

$$(8) \quad var(\bar{I}_2) \;'\; \left[F^2_{2w} \left[\frac{(1-g_c)}{n_{2c}a} + \frac{(1-g_p)}{n_{2p}a} \right] + \frac{(1-f)F^2_{2b}}{a} \right] deff_{2w} \;,$$

where the subscripts *c* and *p* refer to the two groups (for example, the control and program research groups) $deff_{2w}$ is the design effect due to unequal weighting, and where the subscripts denoting dense or nondense areas have been dropped for notational simplicity.

The term F^2_{2b} in equation (8) represents the variance of $\bar{I}$ across areas. In other words, it represents the *extent to which the differences in means vary across areas.* The term captures both the between-area variance in the mean measure as well as the correlation of the group means within areas. The term F^2_{2w} represents the variance of the measure within areas.

An unbiased estimate of the variance expression in equation (8) is as follows:

$$(9) \quad \hat{var}(\bar{I}) = \left[(1-f)\frac{s_b^2}{a} + s_w^2 \left[\frac{f(1-g_c)}{n_c a} + \frac{f(1-g_p)}{n_p a} \right] \right] deff_{2w}$$

where s_b^2 is the sample variance of the difference in the group means between areas, s_w^2 is the (average) sample variance of the measure across youths within areas, and other subscripts are omitted for notational simplicity.

As described in the previous section, it is problematic to estimate the sample variance terms using post-45-day sample members only because of small sample sizes. Thus, we use the *full* within-45-day and post-45-day samples in the selected dense or nondense areas to calculate s_b^2 and s_w^2. We then calculate equation (9) *using post-45 day sample sizes*, and calculate design effects by dividing the estimated variance by an estimate of the variance of the difference between the two means, assuming a simple random sample design.

The design effect for measuring differences in the distributions of the characteristics of control group and program group members is about 1.02. These design effects are small because the differences between the group means is close to zero in all areas. Thus, the design effect for the clustered portion of the sample is less than 1 for most measures.

3. Comparison of the Distributions of Categorical Variables Across Two Groups

In this report and the companion reports, we used a modified chi-squared statistic to test whether the distribution of a categorical variable differs across two groups. This test statistic was constructed by dividing the usual chi-squared statistic (appropriately weighted) by the average design effect across each level of the categorical variable (Scott and Rao 1981). This average design effect was

calculated in two steps. First, we calculated the design effect for comparing the difference between group proportions for *each level* of the categorical variable. The methods from the previous section were used to calculate these design effects. Second, we took a weighted average of these design effects.

Formally, we used the following equations to construct the chi-squared statistic:

$$(10) \qquad P^2_{SR} = \frac{P^2_w}{\bar{d}}$$

$$(11) \qquad P^2_w = \sum_{i=1}^{2} \sum_{j=1}^{J} \frac{(n_i p_{ij} - n_i p_{.j})^2}{n_i p_{.j}}$$

$$(12) \qquad p_{.j} = \frac{n_1 p_{1j} + n_2 p_{2j}}{n_1 + n_2},$$

$$(13) \qquad \bar{d} = \frac{1}{(J-1)} \sum_{j=1}^{J} (1-p_{.j}) d_j ,$$

where p_{ij} is the proportion of youths in group I who are in category j, n_i is the number of youths in group I, $p_{.j}$ is the proportion of the study population in category j, and d_j is the design effect for category j as described above. Under the null hypothesis of no difference between group distributions, the chi squared statistic is distributed chi-squared with (J-1) degrees of freedom.

The modified chi-squared test statistic is intuitive. The statistic decreases as the average design effect increases. Thus, the hypothesis of no difference between group proportions is rejected less often as the average design effect (that is, the average variance across the categories) increases.

REFERENCES

Scott, A.J., and J.N.K. Rao. "Chi-Squared Tests for Contingency Tables with Proportions Estimated from Survey Data." In *Current Topics in Survey Sampling*, edited by D. Krewski, R. Platek, and J.N.K. Rao. New York: Academic Press, 1981.

THIS PAGE INTENTIONALLY BLANK